denodo | BITanium

Architect to Architect Ebook Series:

Data Virtualization for Logical Data Warehouse

Chapter 1

The Logical Data Warehouse

Put simply, the Logical Data Warehouse is a data system that encompasses the concepts of a traditional enterprise data warehouse, while including data from other data sources in addition to that from one (or more) core data warehouses. Though the Logical Data Warehouse embodies something of a technological evolution, it does not serve as a replacement for current data warehouse practices. Instead, the Logical Data Warehouse permits IT organizations to make a large number of data sets and sources available for analysis via query tools and applications, all while improving decision-making and cost savings. There you have it, the "what" and the "why" of the Logical Data Warehouse (also see Figure 1).

Figure 1

The "What" and the "Why"of the Logical Data Warehouse

What is a **LDW**?	A data system that follows the ideas of traditional EDWs, and includes, in addition to one (or more) core DWs, data from external sources. A new data management architecture for analytics combining the strengths of traditional repository warehouses with alternative data management and access strategy.
Why a **LDW**?	Allows the integration of multiple data sources including enterprise systems, the data warehouse, additional processing nodes, web services, cloud, and unstructured data. Permits IT organizations to make large numbers of data sets available for analysis via query tools and applications.

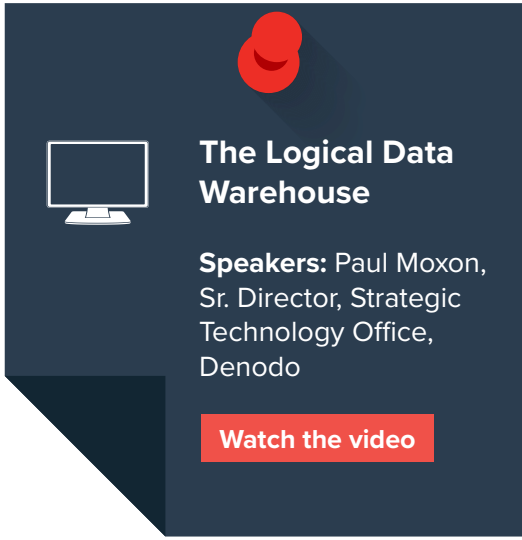
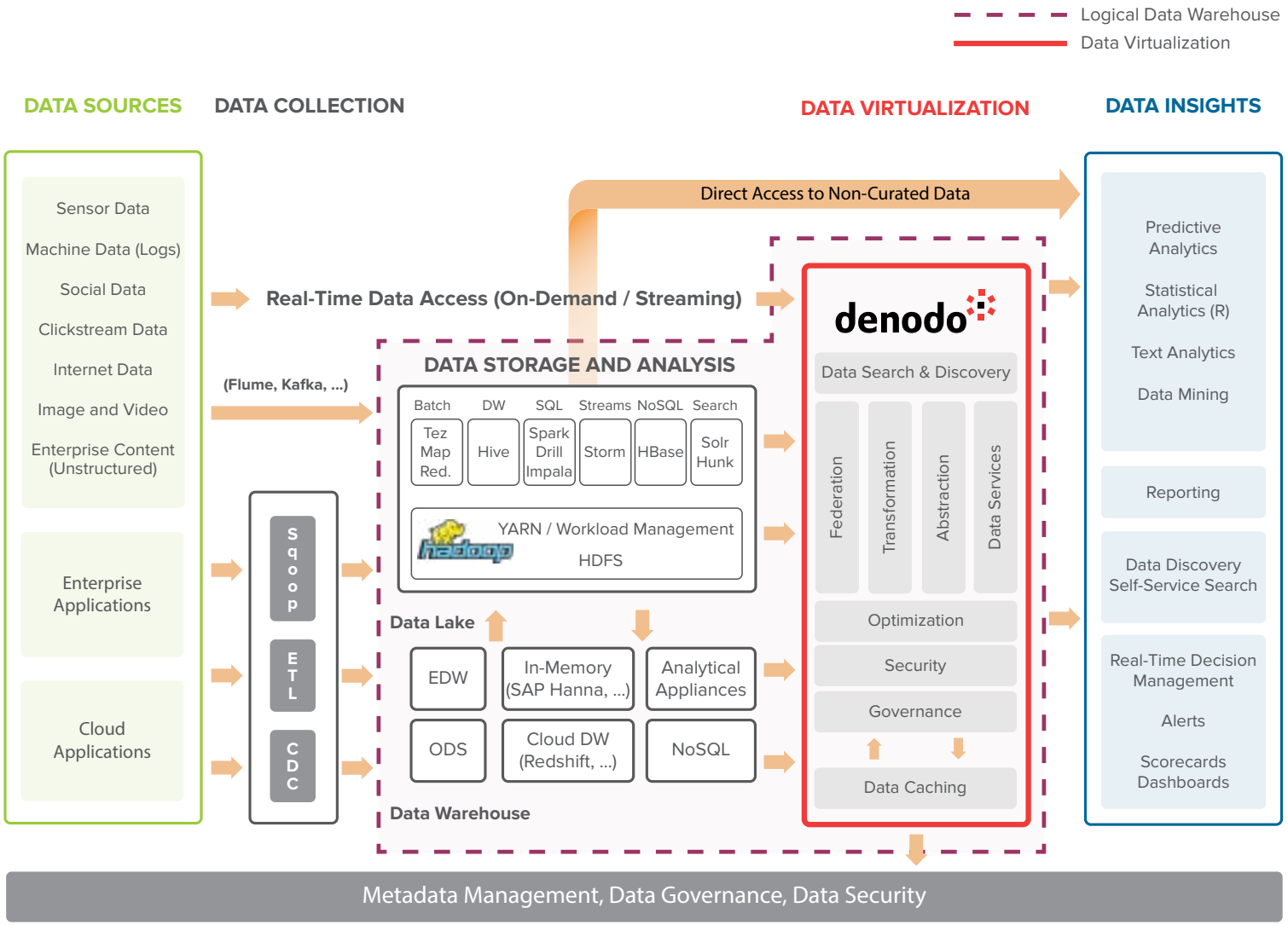


Figure 2 illustrates the Logical Data Warehouse architecture, showing the role of the data virtualization layer acting as an abstraction and semantic layer on top of the underlying heterogeneous data stores. Note the direct access to the data (typically data stored in Hadoop) for non-curated use – not everything needs to go through the data virtualization layer. If users need access to the raw data (the non-curated data) for investigative purposes, etc. then they should not be force to go through the data virtualization layer.

Figure 2

Logical Data Warehouse architecture using Denodo Platform



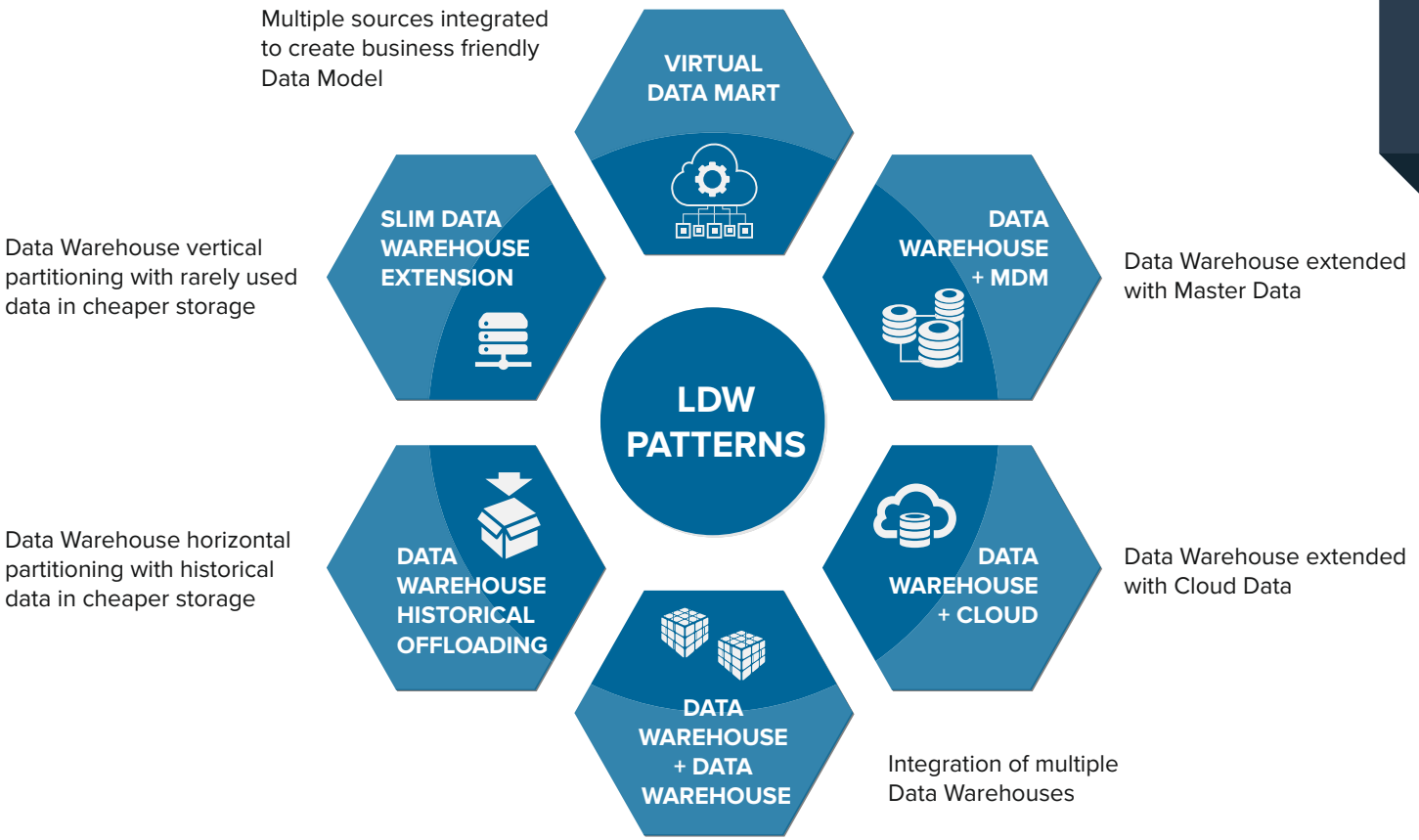
Chapter 2

Common Patterns for Logical Data Warehouse

After numerous implementations of the Logical Data Warehouse, we have discovered six common patterns (Figure 3), each with different fundamental needs, and each delivering a new world of decision making, cost savings, and simplicity. These six patterns are:

Figure 3

Six common patterns of Logical Data Warehouse



Chapter 3

Performance in a Logical Data Warehouse

It's a common assumption that a virtualized solution would be slower than a persisted approach via ETL. However, Denodo has completed extensive testing using queries from the TPC-DS standard benchmarking test for decision support systems, revealing that typical reporting and analytical queries – usually involve aggregation operations over very large data sets – can be executed in a distributed, federated environment, such as the Logical Data Warehouse, with performance that matches that of querying the same data preloaded into a single data source (see Figure 4).

Figure 4

Performance comparisons of Logical Data Warehouse and Physical Data Warehouse

Query Description	Returned Rows	Avg. Time Physical (Netezza)	Time Denodo (Federated Oracle, Netezza & SQL Server)	Optimization Technique (automatically chosen)
		Data pre-loaded into Netezza	Denodo across 3 sources	
Total sales by customer	1.99 M	20.9 sec	21.4 sec	Full aggregation pushdown
Total sales by customer and year between 2000 and 2004	5.51 M	52.3 sec	59.0 sec	Full aggregation pushdown
Total sales by item brand	31.35 K	4.7 sec	5.0 sec	Partial aggregation pushdown
Total sales by item where sales price	17.05 K	3.5 sec	5.2 sec	On the fly data movement

Chapter 3

The queries used are typical reporting/analytical queries – "total sales by customer", "total sales by item for discounted items", etc. These queries scan over very large data sets – 290 million rows in the Sales fact table, 2 million rows in the Customer dimension table, and 400,000 rows in the product dimension table, in the above tests – and return a relatively small results sets by aggregating and summing the raw data.

To achieve these high levels of performance, Denodo has focused on three core concepts:

01

The Dynamic Multi-Source Query Execution Plan leverages the processing power and architecture of data sources. Its "dynamic" nature allows the platform to support ad hoc queries, bringing about its fundamental differentiation, and quite frankly, superiority, to ETL.

02

The Selective Materialization capability intelligently caches only the most relevant and commonly used data, once again, allowing for optimal performance.

03

Finally, the Optimized Resource Management concept enables smart allocation of resources to handle high concurrency. Such resource plans are of course, based on defined rules.

The Dynamic Query Optimizer – depicted in Figure 5 – is a core component of Denodo's approach to a scalable and performant Data Virtualization platform. The Query Optimizer follows a sequence of steps in ensuring that a query runs in the most performant manner possible.

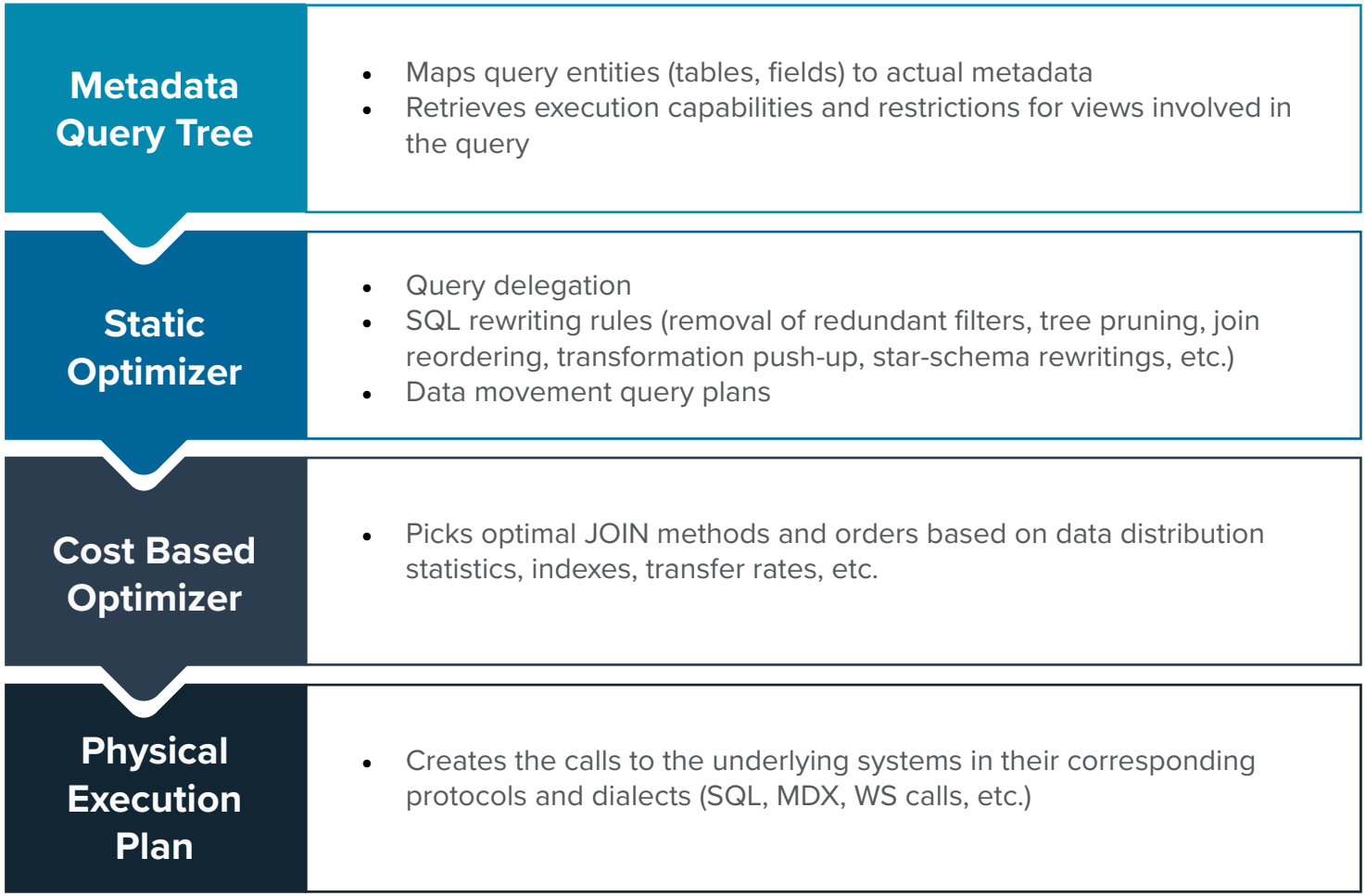
Performance in a Logical Data Warehouse

Speakers: Pablo Álvarez, Principal Technical Account Manager, Denodo

Watch the video

Figure 5

Denodo Platform's Dynamic Query Optimizer Step by Step



Chapter 4

Security and Governance in a Logical Data Warehouse

The wide use of a Logical Data Warehouse architecture – and its natural progression to a Data Services layer for accessing data – means that more and more people within an organization can now access data that they need to perform their daily tasks. This inherently introduces challenges around security and privacy, performance (SLA) and QoS, and, finally, data governance and data veracity. Questions like “How do we secure that data?”, “How do we protect data privacy while still making the data available?”, “How do we guarantee our SLA with more people accessing the data?”, “How do we know where the data comes from...and how ‘good’ is it?”, and so on. These challenges are shown in Figure 6.

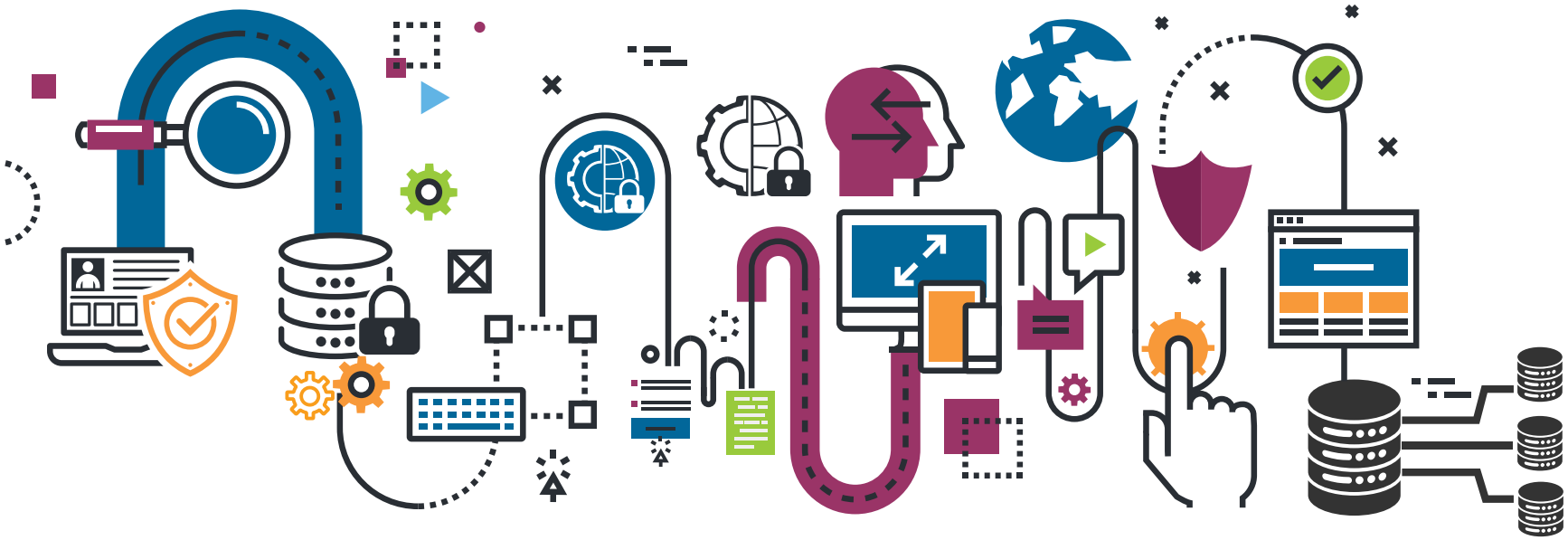


Figure 6

Security and governance challenges in Logical Data Warehouse

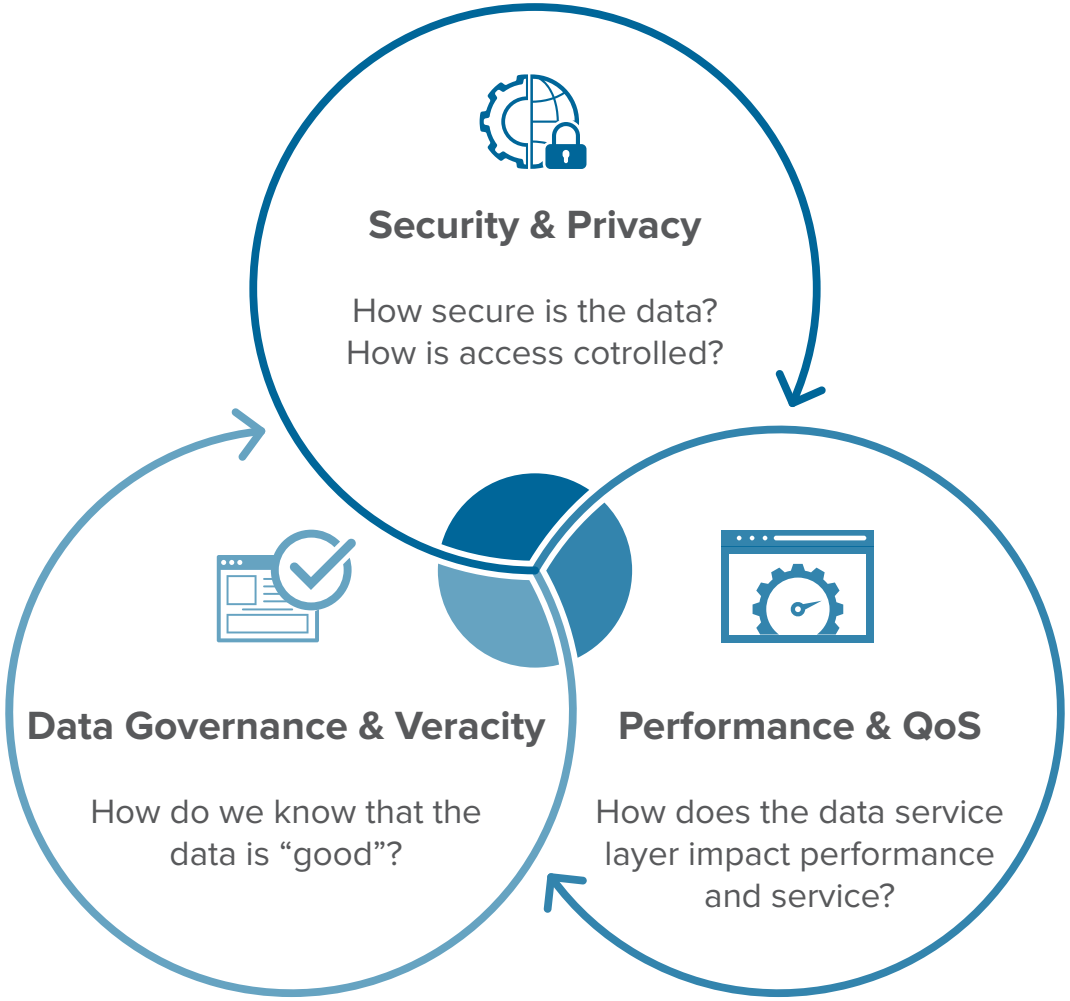
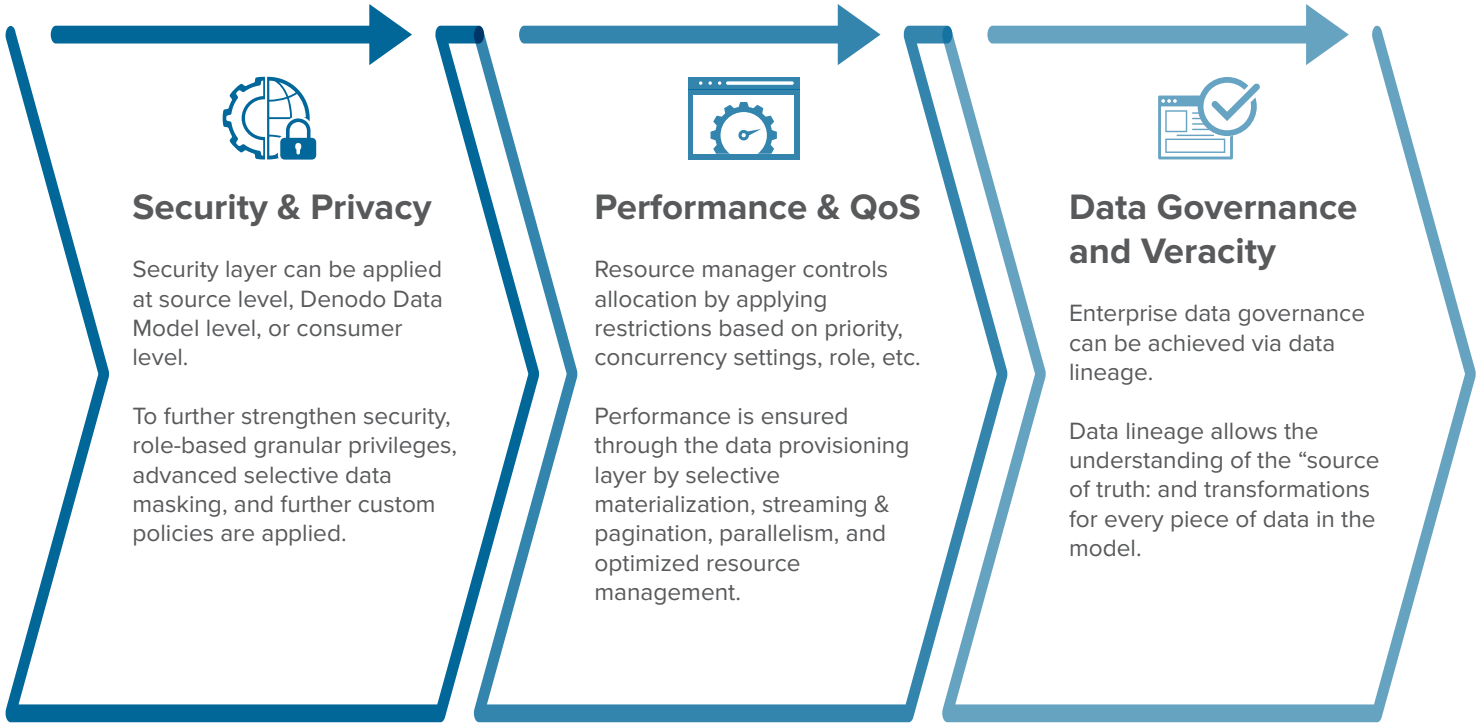


Figure 7

Denodo answers security and governance challenges of Logical Data Warehouse



Security and privacy is critical within a Logical Data Warehouse and Data Services environment. As more users access the data, the more the mechanisms need to ensure security and privacy. This problem is compounded by the fact that the underlying data sources within a Logical Data Warehouse architecture often have different security models – each with different levels of sophistication and maturity. For example, your physical data warehouse (e.g. Teradata, Netezza, Microsoft SSAS, etc.) will have mature security and privacy capabilities. Newer and emerging data technologies, such as Hadoop and Spark, have relatively immature security models and in some cases (e.g. data in HDFS files) the model is "all or nothing" i.e. if you can access the file, you can read all of the data.

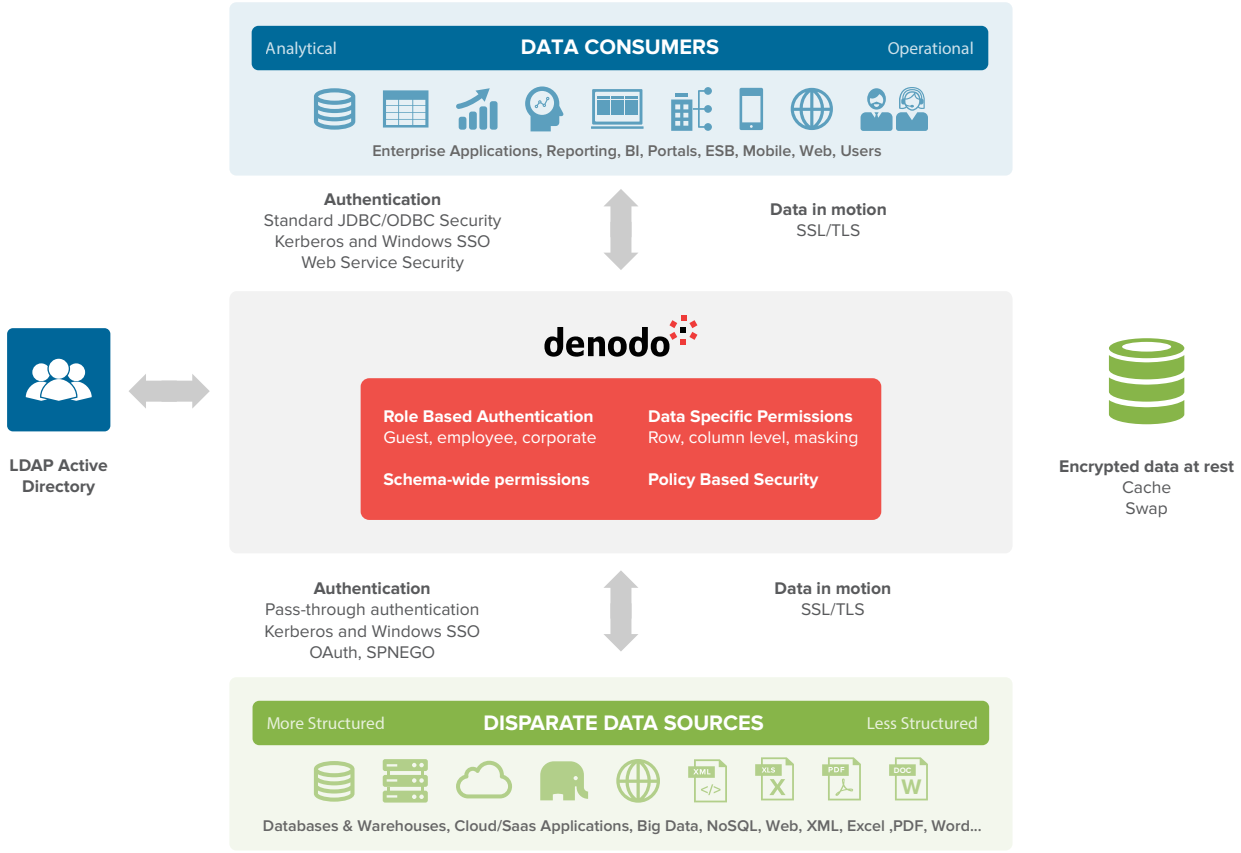
Chapter 4

The Denodo Platform provides a security abstraction and normalization layer which imposes a common – and strong – security model on all of the underlying data sources in the Logical Data Warehouse environment. It provides row, column security, in addition to masking at the value level, irrespective of the data source. So, data in HDFS files can be protected in the same way as data in a mature data warehouse.

Figure 8 provides an overview of the security layers and capabilities when using the Denodo Platform for a data virtualization layer.

Figure 8

Solution view of how Denodo answers security and governance challenges of LDW



Chapter 5

Information Self-Service and Discovery

The advent of the Logical Data Warehouse is also accompanied by more issues than the three mentioned in the previous chapter (i.e. security and privacy, performance and QoS, and data governance and veracity). The Logical Data Warehouse makes more data readily available to people who need that data – and in ways that make it easy for them to consume the data. This, in turn, pushes the notion of 'self-service' for data, allowing users to investigate the data, run simple 'google-like' searches on the data, browse the relationships between data entities, check the data lineage, and build queries on top of the data. This is not intended to replace the final consumer of the data (e.g., a BI/Visualization tool such as Cognos or Tableau), but does allow users to verify that the data is what they want before they start to create their report or dashboard.

The Denodo Platform includes a browser-based Information Self-service tool (ISS) which supports these capabilities and can be used by any user of the data. It doesn't require any special skills or knowledge...just an understanding of what you are looking for in the data. Figure 9 provides a description of the core capabilities included in the ISS tool.

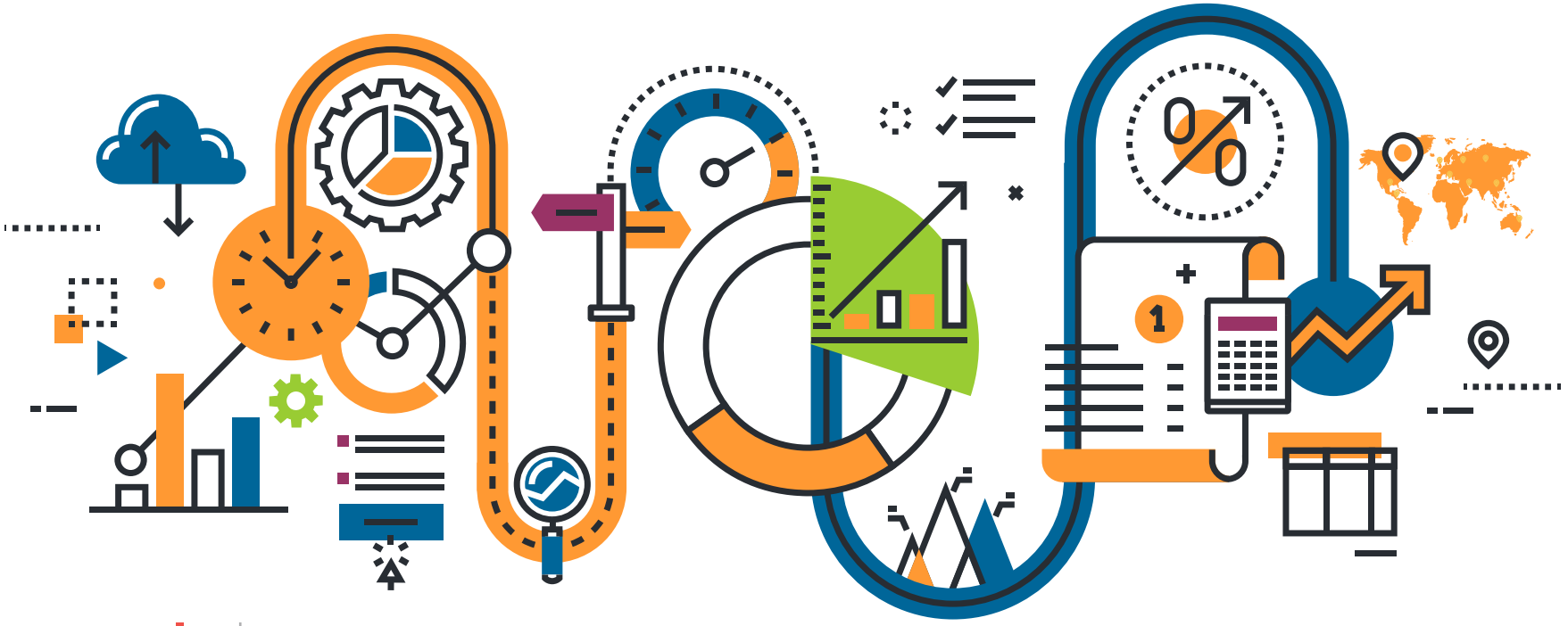
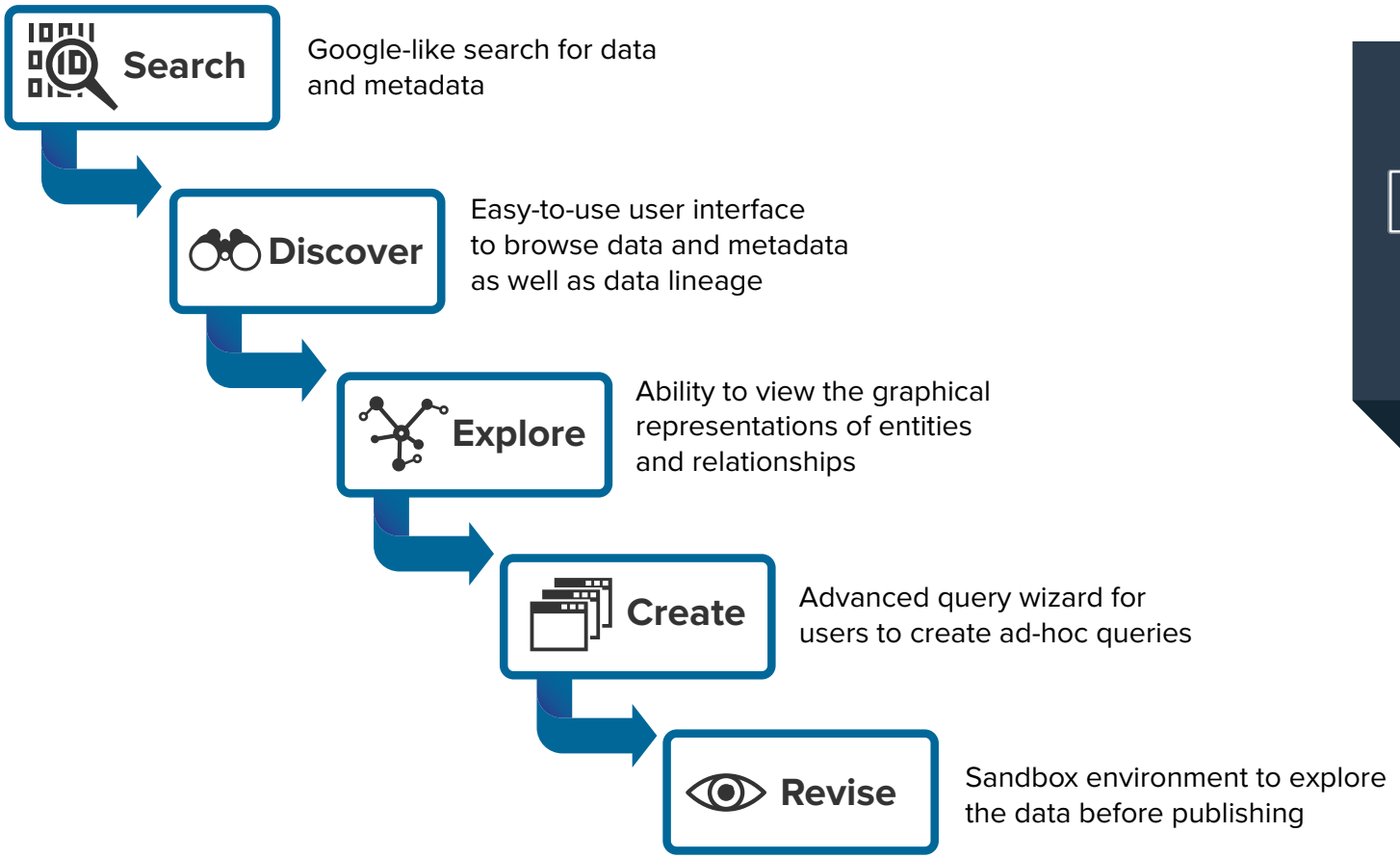


Figure 9

Denodo Platform's Information Self-service and Discovery



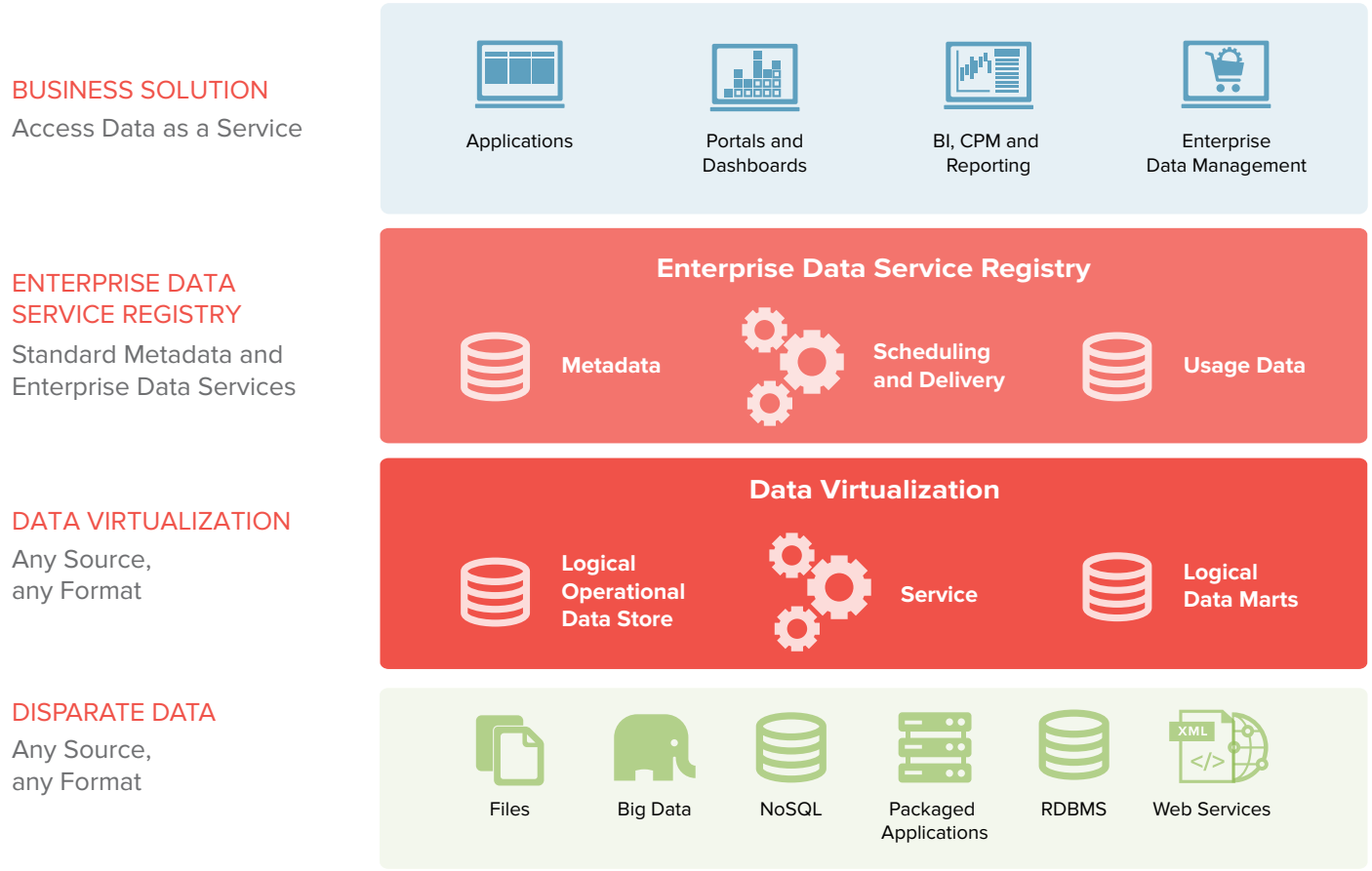
Chapter 6

Building a Data Services Marketplace

A natural extension of building a Logical Data Warehouse – or a Data Services layer – is creating a Data Services Marketplace for the data that has been exposed. This builds a second layer of functionality on top of the Logical Data Warehouse or Data Services layer to allow users to search and browse for data exposed by this layer. This 'enterprise registry' layer provides a more familiar user interface for browsing and requesting access to the data sets offered by the Logical Data Warehouse. Typically, the user interface mimics the 'shopping experience' provided by eCommerce sites such as Amazon. The conceptual architecture for a Data Services Marketplace is shown in Figure 10.

Figure 10

Building a Data Services Marketplace with Denodo



Chapter 7

Logical Data Warehouse Customer Success Stories

The following videos feature Kurt Jackson of Autodesk describing how they used a Logical Data Warehouse – built using the Denodo Platform – to assist in improving to change Autodesk's revenue model to a subscription-based model:

Customer Case Study: Autodesk

Speakers: Kurt Jackson, Platform Architect at Autodesk.

Watch the video

Resources

Case Studies

- Autodesk
- Drillinginfo
- Guardian Life

Videos

- DataNinja Webinar Series
- Educational Seminar (NYC event) videos
- Fast Data Strategy Virtual Event presentations

Whitepaper

- 62-Page Special Report: Managing Data Warehouse Offloading with Big Data

Denodo Recognized as Visionary in the 2016 Gartner Magic Quadrant for Data Integration Tools.

Read this report to understand the vendors who support data virtualization including Denodo, and their strengths and weaknesses.

READ THE REPORT